The Edit Distance as a Measure of Perceived Rhythmic Similarity

OLAF POST
Harvard University, Department of Music

GODFRIED TOUSSAINT
Harvard University, Department of Music

ABSTRACT: The 'edit distance' (or 'Levenshtein distance') measure of distance between two data sets is defined as the minimum number of editing operations – insertions, deletions, and substitutions – that are required to transform one data set to the other (Orpen and Huron, 1992). This measure of distance has been applied frequently and successfully in music information retrieval, but rarely in predicting human perception of distance. In this study, we investigate the effectiveness of the edit distance as a predictor of perceived rhythmic dissimilarity under simple rhythmic alterations. Approaching rhythms as a set of pulses that are either onsets or silences, we study two types of alterations. The first experiment is designed to test the model's accuracy for rhythms that are relatively *similar*; whether rhythmic variations with the same edit distance to a source rhythm are also perceived as relatively similar by human subjects. In addition, we observe whether the salience of an edit operation is affected by its metric placement in the rhythm. Instead of using a rhythm that regularly subdivides a 4/4 meter, our source rhythm is a syncopated 16-pulse rhythm, the *son*. Results show a high correlation between the predictions by the edit distance model and human similarity judgments (r = 0.87); a higher correlation than for the well-known generative theory of tonal music (r = 0.64). In the second experiment, we seek to assess the accuracy of the edit distance model in predicting relatively *dissimilar* rhythms. The stimuli used are random permutations of the son's inter-onset intervals: 3-3-4-2-4. The results again indicate that the edit distance correlates well with the perceived rhythmic dissimilarity judgments of the subjects (r = 0.76). To gain insight in the relationships between the individual rhythms, the results are also presented by means of graphic phylogenetic trees.

Submitted 2011 June 20; accepted 2011 July 17.

KEYWORDS: *edit distance, Levenshtein distance, metrical hierarchies, rhythmic alterations, phylogenetic trees, Mantel test, musical rhythm, music information retrieval, music theory, music perception*

INTRODUCTION

THE 'edit distance' (or 'Levenshtein distance') between two data sets is defined as the minimum number of insertions, deletions, and substitutions of elements in the data, that are required to transform one data set into the other. Though the model is used primarily in data mining and sequence comparison in general (Sankoff & Kruskal, 1999), we seek to assess its value in predicting similarity between two musical rhythms – strings of onsets and silences – as judged by human subjects. (We take the distance between two rhythms, whether calculated or perceived, to be the inverse of the similarity between them.) Two experiments are carried out to test the power of the edit distance in predicting rhythmic similarity. In the first experiment, we test whether rhythms that are all the same edit distance apart from a source rhythm – and twice that distance apart from each other – are also observed by human listeners to be (1) about equally similar to each other and (2) even more similar to the source rhythm. Our source rhythm is the sixteen-pulse *clave son* rhythm, subsequently referred to as 'son.' (See figure 1.) The rhythmic variations on it all differ from it on one pulse: a single onset is substituted for a silence or vice versa. Since all pulse positions are manipulated in a different stimulus, we can also observe whether substitution in certain pulse positions

yields a more salient rhythmic change than in other positions. In the second experiment, we test whether the edit distance model is accurate for rhythmic comparisons for which it predicts different degrees of (dis-) similarity. To generate a set of rhythms with different edit distances, the son is used again; this time by permuting its inter-onset intervals. (The son consists of the inter-onset intervals 3-3-4-2-4.) Listeners are asked to rate the similarity between the rhythms, and the listeners' judgments are compared with the matrix of distances as predicted by the edit distance model. Note that in the first experiment, the number of onsets within the set of rhythms varies from 4 to 6, and that in the second experiment this number is the same for all the stimuli, namely 5. The results of the two experiments are analyzed not only statistically, but also graphically, using phylogenetic trees. (See figures 5-7, 11, and 13.) While the distances in the phylogenetic trees do not always represent distance relationships exactly, they chart the distance of each individual rhythm to the others, thus offering direct insight in possible clusters and outliers among the rhythms.

Previous studies on the perceptual effects of rhythmic alterations are few. In one noteworthy study Patricia Sink showed that alterations such as augmentation, diminution, retrograde, retrograde-diminution, retrograde-augmentation, repeating rhythmic values, and the addition of a melodic pattern altered subjects' perception of rhythmic patterns (Sink, 1983, 1984). However, no previous studies have been reported on the alterations used here. Hannon and Trehub (2005) studied the perceptual effects of alterations that either preserved or violated the original metric structure of the rhythms. In our experiments, all of the rhythms have the same number of pulses, though no effort is made to guarantee either preservation or violation of the metrical structure. After all, one does not need to determine a meter for the rhythms that one is comparing in order to use the edit distance model; what the edit distance model does is simply calculate the smallest number of manipulations required to transform one rhythm into another.

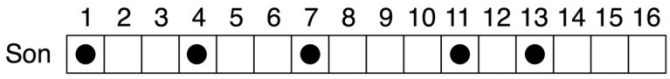

Fig. 1. The 16-pulse clave son rhythm (in box notation)

PREDICTIVE MODELS OF RHYTHMIC SIMILARITY

The Edit Distance Model

Unlike the study of rhythmic similarity with acoustic input (e.g., Smith, 2010), the models used in this paper assume that the rhythms are represented symbolically. Specifically, the models assume that a rhythm is represented as a binary sequence of unit time pulses that are either sounded (onsets) or silent (rests). Thus rhythms can be conveniently expressed as binary sequences of symbols. For example, the 16-pulse son can be notated as [x . . x . . x . . . x . x . . .], where 'x' is the symbol denoting a sounded pulse, and '.' the symbol denoting a silent pulse. In the context of the edit distance model, a deletion of either an onset or a rest converts the son to a 15-pulse rhythm. Similarly, an insertion of either a rest or an onset converts the son to a 17-pulse rhythm. A substitution of an onset for a rest or a rest for an onset does not change the number of pulses of the rhythm. For example, the son can be converted to the rhythm [x . . x . . x . . . x . . . . .] by either substituting the fifth onset by a rest, or by first deleting the fifth onset, and then inserting a rest in its location. In this example, the substitution of the fifth onset yields the smallest edit distance, 1, whereas the deletion followed by an insertion yields an edit distance of 2. In rhythmic manipulations where the length of the rhythms remains constant, substitutions often make for more efficient steps than insertions or deletions (Orpen and Huron, 1992). In addition to the edit distance model, we are testing two other models for the prediction of rhythmic similarity: an approach based on Lerdahl and Jackendoff's Generative Theory of Tonal Music (GTTM; Lerdahl & Jackendoff, 1983) and the swap distance model (Toussaint, 2004).

Generative Theory of Tonal Music

For a 16-pulse rhythm, Lerdahl and Jackendoff (1983) propose a rhythmic hierarchy of stronger and weaker pulses that is illustrated in figure 2. The heaviest weight is assigned to pulse 1, the second heaviest at pulse 9, the third at pulses 5 and 13, the fourth at positions 3, 7, 11, and 15. The remaining eight pulses at positions 2, 4, 6, 8, 10, 12, 14, and 16 all receive the minimum weight. (Note that the numbers on the y-axis represent the ordinal weight only; no specific units are implied in these numbers.) Drawing on Huron's *Sweet Anticipation* (Huron, 2006), one could also say that the metric weight of a pulse represents the general likelihood of an onset on that pulse. Thus rhythms containing onsets at positions with low metric weights would tend to be more complex or syncopated. Rhythmic similarity could then be quantified on the basis of the degree and distribution of their syncopation.[1]

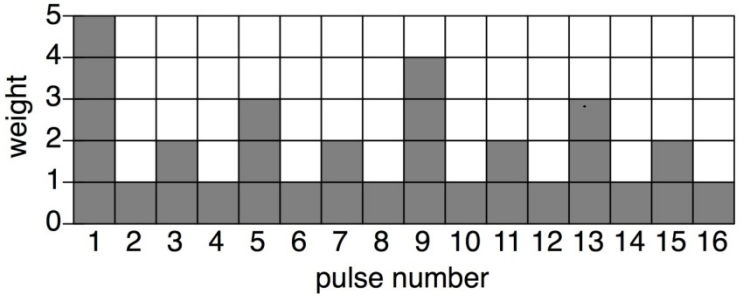

Fig. 2. The metric weights assigned to the sixteen positions of a 16-pulse rhythm according to GTTM

The Swap Distance Model

Another rhythmic similarity measure that has been applied with some success in previous comparative studies (Toussaint, 2004) is the swap distance model. A *swap* is the position interchange of an onset and an adjacent silent pulse. (Swapping two similar pulses would have no effect.) The swap distance between two rhythms is the minimum number of swaps required to convert one rhythm to the other. When the two rhythms being compared have a different number of onsets (as is the case in Experiment I), two constraints are added to the above definition: (1) Every onset of the first rhythm must move to the position of an onset of the second rhythm, and (2) Every onset position of the second rhythm must receive at least one onset from the first rhythm. For example, consider the first and third rhythms in figure 3 below: the son and D-4. The son is the denser rhythm, since it has five onsets; D-1 is the sparser rhythm since it has four onsets. Without the first constraint, onsets 1, 7, 11, and 13 of the son could be mapped to onsets 1, 7, 11, and 13 of D-4 at a cost of zero swaps, but the disappearance of onset 4 of the son would not be considered for the distance calculation. However, constraint (1) forces onset 4 of the son to be assigned to an onset of D-4. Since it takes three swaps to move onset 4 to either onset 1 or 7, the swap distance between the rhythms is three. Now consider the fourth and the fifth rhythms in figure 3: rhythms D-7 and D-11. Without the second constraint, onsets 1, 4, and 13 of D-7 would be mapped to onsets 1, 4, and 13 of D-11, and onset 11 of D-7 would also be mapped to onset 13 of D-11, for a total cost of 2 swaps. However, onset 7 of D-11 would not be considered in the comparison. Constraint (2) forces onset 11 of D-7 to be assigned to onset 7 of D-11, leading to a swap distance of four instead of two between the rhythms.

EXPERIMENT I: PREDICTING RHYTHMIC SIMILARITY

Rhythms Used

Besides the son itself, sixteen rhythms were used: all possible single-onset changes of the son. Five of these rhythms are the result of substituting an onset for a silence; 11 are the result of substituting a silence for an onset. The first group we refer to as 'deletions' – though in terms of the edit distance model they are not

deletions but substitutions. They are labeled with the letter D followed by the number of the pulse on which the substitution occurs (i.e., D-1 is the rhythmic derivation of the son in which the first onset is substituted for a silence). See figure 3. The second group we refer to as 'insertions.' They are labeled with the letter I followed by the number of the pulse in which a silence is substituted for an onset. See figure 4.

DELETIONS

Figure 3 shows, in box notation, the son rhythm (top) and the five altered rhythms obtained by replacing each of its five onsets with a silent pulse.

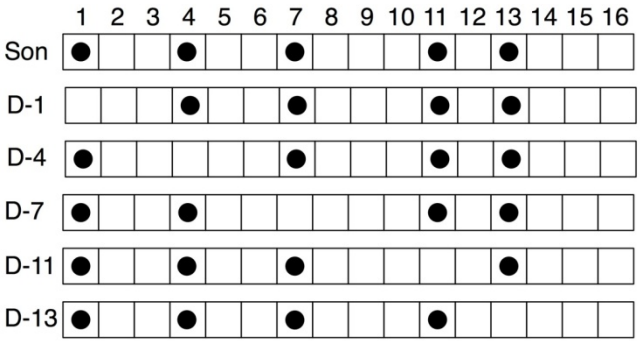

Fig. 3. The son (top) and the five rhythms obtained by substituting each of its onsets for a silence.

INSERTIONS

Figure 4 shows the son rhythm (top) and the eleven altered rhythms obtained by replacing each of its silent pulses with an onset.

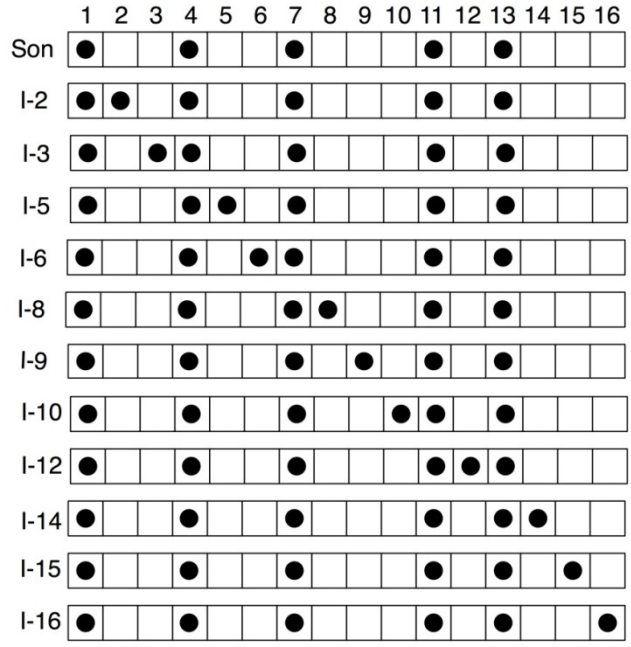

Fig. 4: The son (top) and the insertions; the eleven rhythms obtained by replacing one of the eleven available silences with an onset.

PHYLOGENETIC TREES OF THE PREDICTED DISTANCES

In Figure 5, the edit distance between the son and the rhythmic variants – the deletions and insertions – is represented visually by means of a phylogenetic tree. The tree is calculated using the matrix of edit distances between the 17 rhythms, and it is drawn using the program *BioNJ* (Gascuel, 1997) embedded in the *SplitsTree-4* software package (Huson, 1998). This program draws the tree such that, if possible, the 'linear distance,' the shortest path along branches in the tree between any pair of rhythms, corresponds to the edit distance in the distance matrix. Predictably, the visual distance in the tree between the son and each of the altered rhythms is the same (corresponding to an edit distance of 1), and the visual distance between all other pairs of rhythms is twice as long (corresponding to an edit distance of 2).

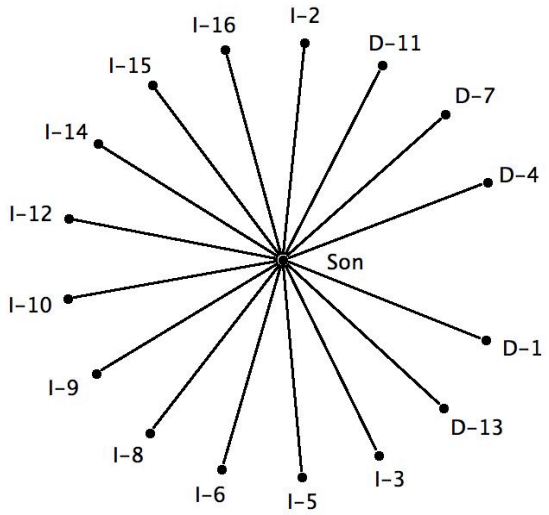

Fig. 5: Phylogenetic tree of the edit distance between the 17 rhythms computed with the edit distance matrix. The visual distance between the son and all of its variations represents an edit distance of 1; that between the sixteen variations themselves an edit distance of 2.

The phylogenetic tree of the rhythms calculated with the swap distance is shown in Figure 6.

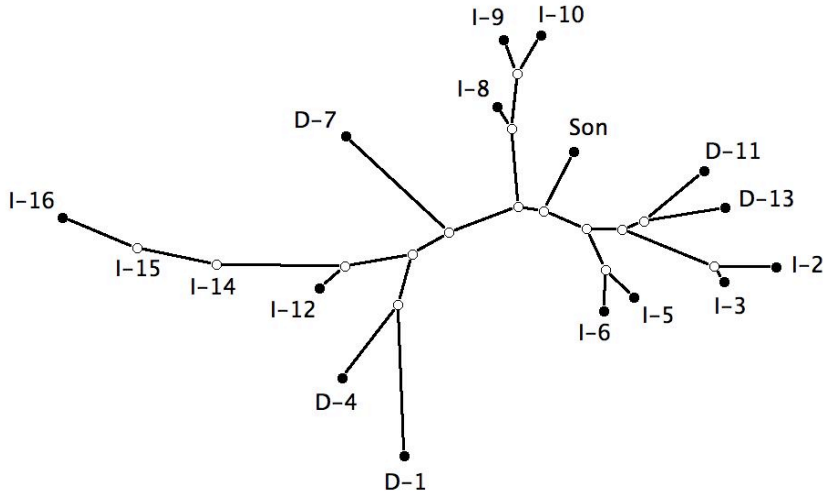

Fig. 6: Phylogenetic tree of the 17 rhythms computed with the swap distance matrix.

A comparison of the two phylogenetic trees (Figures 5 and 6) suggests that the edit distance model and swap distance model calculate quite different distance relations between the rhythms. In contrast to the edit distance model, the distance values in the swap distance model of the I-rhythms and D-rhythms to the son vary widely. (Recall that in the edit distance model this distance is 1 for all of the I- and D-rhythms.) For example, the son is relatively close to I-5, I-6, D-11 and D-13, but relatively distant from I-15, I-16, D-4, and D-1. However, in observing these phylogenetic trees, it should be noted that the visual distances are not always exact, as the complex sum of all of the distances defined in the matrix cannot always be represented accurately in two dimensions. If that is the case, the plotted phylogenetic tree is an approximation in which the total discrepancy between the tree and the distance matrix is minimized.

While these visual representations give a good overview of the distances and clusters among the rhythms, their use in establishing the exact degree of similarity between the distance matrices is limited. A more precise way to determine the similarity between the matrices is a statistical Mantel test (Bonnet & Van de Peer, 2002), with which a correlation coefficient between two matrices can be calculated.[2] For the two distance matrices of the 17 rhythms, a one-tailed Mantel test yields a somewhat low correlation coefficient ($r = 0.379$, $p = 0.063$).

Participants

As discussed above, the aim of this experiment is to compare the similarity between the 17 rhythms as calculated by the edit distance model and the swap distance model to similarity judgments made by human subjects. The listening experiment involved a total of 10 participants, 5 females and 5 males (mean age = 22, range = 18-33). The subjects were graduate and undergraduate students from music and engineering programs at Harvard University, who were paid $25 for their participation. The average number of years of musical training among all participants was 13, mainly in classical music.

Apparatus

The subjects were seated on a comfortable swivel chair and listened to the rhythms using Able Planet, model NC1000 noise-cancelling headphones. The headphones were connected to a MacBook Pro laptop Apple computer placed in front of the subjects. As interface we used *Sonic Mapper* software (developed by Gary P. Scavone), which is a combination of *Qt* for the visual interface and *RtAudio* for audio output (Scavone, Lakatos, & Harbke, 2002). Of the three sound comparison procedures available in *Sonic Mapper* – two-dimensional similarity mapping, sorting, and pairwise comparison tests – we used the pairwise comparison procedure.

Stimulus Materials

The sound samples of the rhythms were created in Apple's *GarageBand* software. The rhythms were artificially synthesized and the durations were exact. Each onset was represented by the same identical high-pitched click: the sound of the wooden claves as available in *GarageBand*. Within each stimulus, the specific rhythm was played three times in succession at a tempo of 100 beats (or 400 pulses) per minute; between repetitions of the rhythm a brief pause of 5 pulses was included. Altogether each sound sample lasted for approximately 9 seconds.

Procedure

The experiment took place in a quiet room. Before the start of the experiment, the participant filled out a consent form and a brief biographical data form. The participants were told that they would be hearing pairs of rhythms, and that their task was to rate how similar these rhythms *felt* to them. This they could indicate on a sliding scale from 'least similar' to 'most similar.' (The program then assigned a numerical value to their judgment, ranging from 1 for 'least similar' to 9 for 'most similar.') To familiarize the subjects with the procedure, they were asked to rate four randomly selected pairs of stimuli prior to the experiment. In the actual experiment, the subjects rated all of the 136 possible pairs of stimuli. (The order of all the comparisons was determined randomly, and within each comparison, the order of the two stimuli was determined randomly.) Between the stimuli, there was a pause of 3 seconds. The subjects had the option of

replaying the comparison as often as they wished. The subjects were free to take breaks. Altogether most subjects completed the task in 60 to 75 minutes. After the experiment, the subjects were debriefed and were given a short questionnaire that asked them to identify any strategies they had used in completing the task.

RESULTS OF EXPERIMENT I

To analyze the results of the similarity judgments between the 17 rhythms, we averaged the judgments that were made by all 10 participants. The phylogenetic tree of the perceptual distance judgments is shown in Figure 7. Comparing this tree with the trees for the edit distance model and the swap distance model (Figures 5 and 6), it appears that the edit distance model is a more accurate representation of the human judgments than the swap distance model. One-tailed Mantel correlation tests between the human distance judgment matrix and the predictive matrices are consistent with this hypothesis: The correlation coefficient found for the perceptual matrix and the edit distance matrix is $r = 0.866$ with a probability value of $p = 0.0001$, whereas for the swap distance and the human judgments a correlation coefficient was found of $r = 0.17$ with $p = 0.02$. (Note that the Mantel test yields no degrees of freedom.)

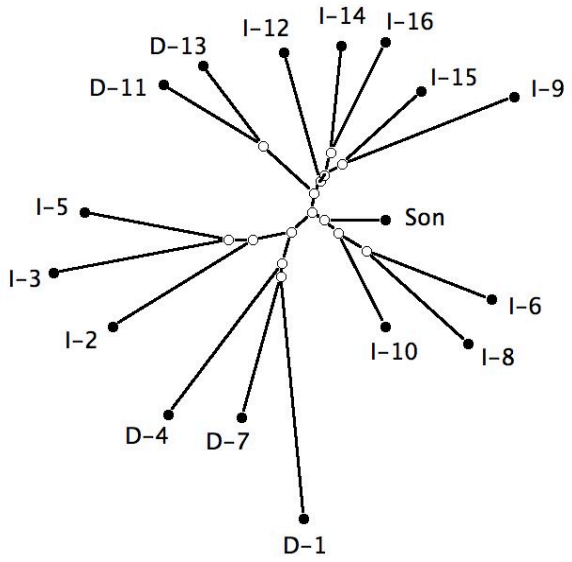

Fig. 7: Phylogenetic tree of the distance matrix of the human rhythmic similarity judgments.

To look at these results in a different way, recall that the edit distance between the son and its altered versions is 1, and that the distance between all of the altered versions is 2. (See Figure 5.) Following this model, we hypothesized the following: for human listeners, the son should be relatively similar to all of the altered rhythms, and all of the altered rhythms should be less similar to each other than to the son. In Figure 8, the human similarity ratings are plotted for these two separate categories: on the left the average similarity rating between the son and the altered rhythms, and on the right the average similarity rating between all of the altered rhythms alone. As predicted, the listeners judged the altered rhythms as more similar to the son ($\mu = 6.7$, $\sigma = 0.8$) than to each other ($\mu = 5.3$, $\sigma = 0.9$). A t-test on this data yielded $t = 6.325$; df = 15; $p = 0.00001$.

Unlike the edit distance model, the swap distance model does not make such a clear-cut difference between distances that involve the son and distances between altered rhythms alone. Although the average swap distance between the son and altered rhythms is indeed smaller than that between the altered rhythms alone ($\mu = 2.0$ versus $\mu = 5.0$), the standard deviation for the second group is much bigger ($\sigma = 1.0$ versus $\sigma = 3.0$), thus allowing for considerable overlap between the two categories. Since these categories seem quite distinct in the results of the human listening test, again this analysis suggests that the edit distance model is a better predictor of perceived rhythmic similarity than the swap distance model, at least under these experimental conditions.

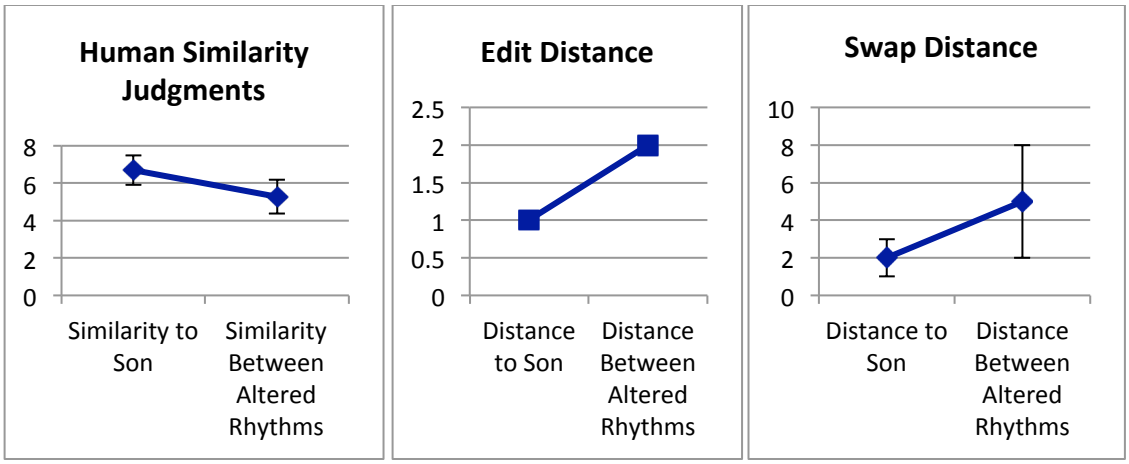

Fig. 8: Comparing the edit and swap distance similarity predictions against human judgments. In the first graph, the vertical axis represents human distance ratings; in the second and third graphs the vertical axis represents the calculated edit and swap distances, respectively.

In order to compare the results in a meaningful way to the predictions that GTTM makes for a 16-pulse rhythm, the salience of each pulse needs to be extracted from the experimental data. A possible approximation of each pulse's salience is to look at the similarity rating between the son and each of its 16 variations, and invert the rating. Thus we obtain a dissimilarity (or distance) rating. For each pulse manipulation, whether deletion or insertion, the dissimilarity rating can be taken to indicate the salience of this manipulation. (A more specific analysis of the salience of the insertions versus the salience of the deletions follows.) In Figure 9, the calculated salience of each pulse is listed along with its metric strength according to GTTM, and the data is plotted in a bar chart. Recall that the five deletions occur on pulses 1, 4, 7, 11, and 13, and the insertions occur on pulses 2, 3, 5, 6, 8, 9, 10, 12, 14, 15, and 16. In order to avoid using the exact numerical values given by GTTM, a ranked statistical correlation test was performed. Between the approximated salience of each pulse and the metric strengths proposed by GTTM, a Spearman rank-correlation coefficient was found of $\rho = 0.64$ with $p = 0.0008$. These results suggest a correlation between the human similarity judgments and the metric placement of the rhythmic manipulations. They also lend support to the conclusions made by Palmer and Krumhansl (1990) on the mental representation of musical meter as being composed of a hierarchy of inferred accent strengths.

Mutation	Human Score	GTTM-Hierarchy
D-1	5.128	5
I-2	4.184	1
I-3	3.288	2
D-4	3.72	1
I-5	3.168	3
I-6	2.6	1
D-7	3.736	2
I-8	2.784	1
I-9	4.592	4
I-10	2.464	1
D-11	3.408	2
I-12	2.552	1
D-13	2.816	3
I-14	2.592	1
I-15	3.112	2
I-16	2.72	1

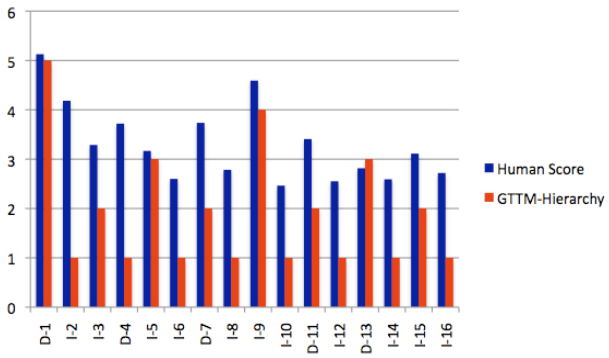

Fig. 9: The salience of each pulse manipulation compared with the GTTM hierarchy values.

D-rhythms versus I-rhythms

In comparison to other popular sixteen-pulse rhythms with five onsets, it has been argued that the son functions as a prototype.[3] If this is indeed the case, one would expect each of these five onsets to be necessary – and the set together to be minimally sufficient – to make the rhythm effective and recognizable. This, in turn, suggests the hypothesis that a deletion of an onset results in a more drastic change to the son than the addition of an onset, the latter being in effect more of an 'unnecessary' decoration. The phylogenetic tree in Figure 7 seems to support this hypothesis, as do the bar charts in Figure 10. To test this hypothesis, the data was analyzed separately for the rhythmic variations with a deletion (the D-rhythms) and with an insertion (the I-rhythms). The average distance from the son to its deletions is $\mu = 3.76$ with $\sigma = 0.85$, and to its insertions is $\mu = 3.096$ with $\sigma = 0.7$. This difference approaches significance; an unpaired *t*-test comparing the two means yields a *p*-value of 0.12.

We can also compare the D-rhythms and the I-rhythms using their separate distance matrices and phylogenetic trees. Figure 11 shows the separate phylogenetic trees calculated for the D-rhythms (left) and I-rhythms (right). Note the place of the son rhythm in the trees: the structures suggest that in the graph of the I-rhythms, the son takes a less central place, indicating greater individual difference between the salience of the insertions. Accordingly, we can hypothesize that the edit distance model is a better predictor for the D-rhythms than for the I-rhythms. Mantel tests of the data are consistent with this hypothesis. The Mantel correlation coefficient found between the edit distance model and the human judgment matrices for the D-rhythms is $r = 0.92$ ($p = 0.003$), whereas for the I-rhythms it is $r = 0.785$ ($p = 0.0001$).

Fig. 10: salience ratings of the I-rhythms (left) and the D-rhythms (right).

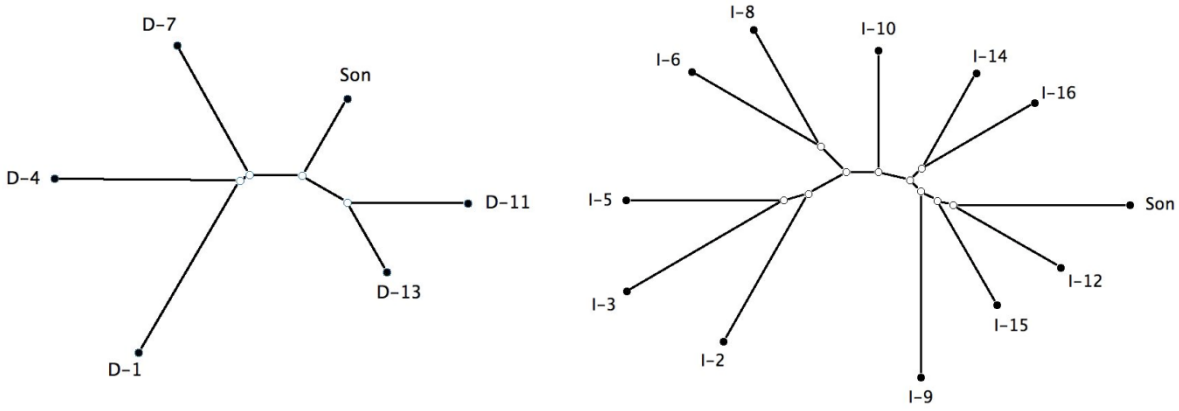

Fig. 11: The phylogenetic trees for the deletions (left) and insertions (right).

EXPERIMENT II: PREDICTING RHYTHMIC DISSIMILARITY

In Experiment I, we used rhythms that – in terms of the edit distance model – are all highly similar: the son and the 16 possible rhythms that are only one substitution away. However, the rhythms varied in two respects that in our opinion warranted a second experiment: (1) the number of onsets present in the rhythms, and (2) their inter-onset-interval (IOI) content. (1) Recall that in experiment I, 5 rhythms were the result of substituting an onset for a silence, reducing the number of onsets to 4. In 11 cases, a silence was substituted for an onset, increasing the number of onsets to 6. Thus the experiment included rhythms with 4, 5, or 6 onsets. This might have been an important factor in the subjects' similarity judgments. (2) Within rhythms with the same numbers of onsets, such as I-8 and I-9, the inter-onset interval structure (i.e., the specific lengths of the five durations that fill the 16 pulses) often varies. For example, I-8 contains one interval of duration 2, whereas I-9 contains three such intervals. Again, these variations in IOI-content may have informed the human similarity judgments in experiment I. In Experiment II, we aim to observe how the edit distance model compares to human judgments in the absence of these two factors. Starting from the son, we randomize the order of its inter-onset intervals, thus obtaining rhythms with the same number of onsets and the same IOI-content. However, in terms of the edit distance model, this set of rhythms is far more varied than the rhythms of experiment I, including edit distances that vary from 2 to 5. (To us, some of these recombinant rhythms were indeed surprisingly tricky to tap, such as 43243). Hence, a secondary objective in the second experiment is to test the edit distance model in predicting different degrees of perceived rhythmic complexity.

Rhythms Used

The stimuli used in Experiment II include the son and permutations of its inter-onset intervals (33424). Theoretically, 30 different rhythms can be made with the IOI-content of the son. However, we believed that using 30 stimuli would be unfeasible due to time constraints. In fact, according to participant feedback, the listening task of Experiment I was experienced as rather long. Therefore, for Experiment II we reduced the number of stimuli, and only used the son and twelve of its permutations. The permutations were ranked according to their edit distance from the son, and from each distance category we randomly selected 3 permutations. In figure 12, the 13 rhythms are shown in lexicographical order of their IOI-structure. Note that a number of these permutations are merely rotations of one another; the IOI-order is identical, but the rhythm starts at another duration (for example, 24433 and 44332). For this reason, we chose not to repeat the rhythms within each stimulus, as this presentation method would most likely affect the similarity judgments. (See the section on stimulus materials below.)

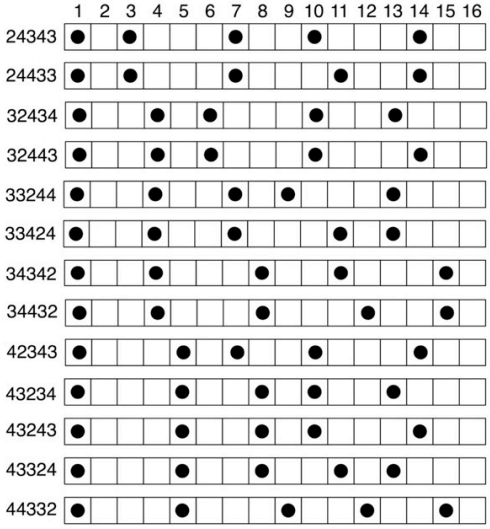

Fig. 12: The 12 selected permutations of the inter-onset intervals 33424 and the son, in lexicographical order.

Participants

The listening experiment involved a total of 10 participants comprising 8 females and 2 males (mean age = 21, range = 18-27). The subjects were undergraduate and graduate students mainly in social sciences, applied mathematics, and life sciences, who were each paid $20 for their participation. The average number of years of musical training among all participants was 8, mainly in classical music.

Apparatus

The apparatus used in Experiment II was the same as used in Experiment I (see above).

Stimulus Materials

The sound samples used in Experiment II were created using Apple *GarageBand*. Again, the rhythms were artificially synthesized and the durations were exact. Each onset was represented by the same identical high-pitched click: the sound of the wooden claves as available in *GarageBand*. However, unlike Experiment I, the rhythm was not repeated within each stimulus; instead, each rhythm was played only once at a tempo of 100 beats per minute (or 400 pulses per minute), resulting in a sound sample that lasted for approximately 2 seconds.

Procedure

The procedure used was the same as that in Experiment I (see above), although fewer similarity judgments needed to be made (78 instead of 136). Since the rhythms were not repeated within the stimuli, and since the rhythms were often relatively complex, we assumed that the participants would listen to each pair of stimuli more often than the participants did in the first experiment. That assumption turned out to be incorrect; more comparisons were heard only once in the second experiment than in the first experiment (1,243/1,360 = 90% for Experiment II versus 668/780 = 85% for Experiment I). However, certain specific comparisons in Experiment II may indeed have been considered relatively difficult compared to Experiment I. For example, in Experiment I, no comparison was replayed three times or more, whereas in Experiment II this happened in three instances.

RESULTS OF EXPERIMENT II

The phylogenetic trees of the son and its 12 permutations are shown in Figure 13. The tree on the left represents the distances as predicted by the edit distance model; the tree on the right represents the distances as judged by the subjects in the listening test. The numbers indicate the IOI-structure of the stimuli. Visually, there are a number of similarities between the trees: in both trees, the son is close to 43324 and 33244, and far removed from 42343. Also, 24343 and 24433 are grouped together in both trees. However, differences can also be observed. For example, 32434 stands by itself in the edit distance model (relatively close to the son), but is grouped with 32443 and 42343 in the tree representing the perceived distances (on the far right). Analyzing the distance matrices statistically, a one-tailed Mantel test yields again a high correlation coefficient: $r = 0.828$, with $p = 0.0001$. This correlation coefficient suggests that the edit distance model is a fairly robust predictor of the perceptual distance between two rhythms, even when participants cannot merely rely on the number of onsets, and when rhythms are relatively dissimilar and difficult to analyze metrically.

Fig. 13: The phylogenetic trees for the edit distance (left) and the perceptual distance (right).

An alternative method for evaluating the effectiveness of the edit distance model in predicting perceptual dissimilarity is to conduct a regression analysis between the perceived distance to the son and the edit distance to the son. Since the distance ratings by the subjects are scalar data, we simply ranked the perceived distances from low to high, and ranked the edit distances in the same order (see Figure 14). With these data, a ranked Spearman correlation coefficient was calculated of $\rho = 0.755$ with $p = 0.004$. Again, the correlation coefficient found is high; as shown in Figure 14, for the first nine rhythms the predictions of the edit distance model are in line with the distance judgments made by the subjects. However, for the rhythms rated most dissimilar to the son, the corresponding edit distance is actually lower than one would expect, especially in the case of rhythm number 10, 32443 (The list is given in Figure 15.)

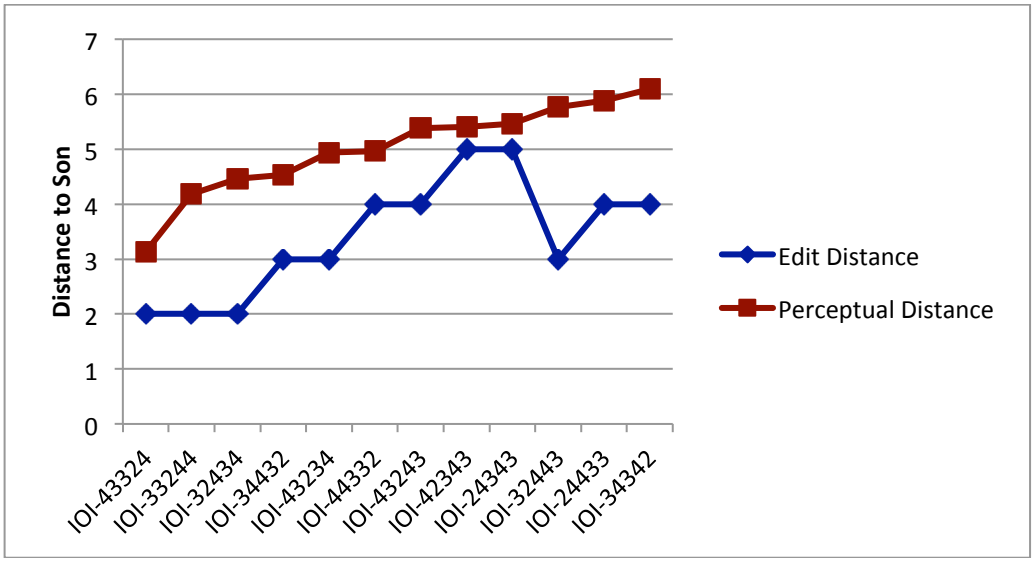

Fig. 14: The perceptual distances (top) plotted by increasing distance to the son, along with their corresponding edit distances (bottom), for each IOI permutation of the son (indicated at bottom).

To gain insight in the discrepancy between the edit distance model and the subjects' distance ratings, the rhythms are displayed in box notation in Figure 15, ranked in order of perceptual distance from

the son from lowest to highest. In the left column, the five pulse positions of the son are highlighted, and in the list on the right the four quarter-note positions are highlighted. The number on the left of each rhythm indicates the inter-onset interval vector of the rhythm, and the numbers on the right are the distances from the son. The first number is the perceptual distance and the number in parentheses is the corresponding edit distance. While a detailed analysis of the differences between the perceived distance and the edit distance does not seem justified on the modest data set of this experiment, the figure may suggest several patterns.

To start with the left column (in which the son rhythm is highlighted), the rhythm that is perceived as most similar to the son shares its last two pulses with it. The subsequent two rhythms, share their second and last onsets with the son. The next rhythm, 34432, only shares its second onset with the son. Most of the remaining rhythms do not share any onsets with the son, with the exception of the last rhythm, 34342. This rhythm, while having its first two onsets in common with the son, has a very late last onset on pulse 15. Thus, it appears that the more characteristic features of the son – in this experimental context – are its last onset on pulse 13, its fourth onset on pulse 11, and to a slightly lesser degree its second onset on pulse 4.

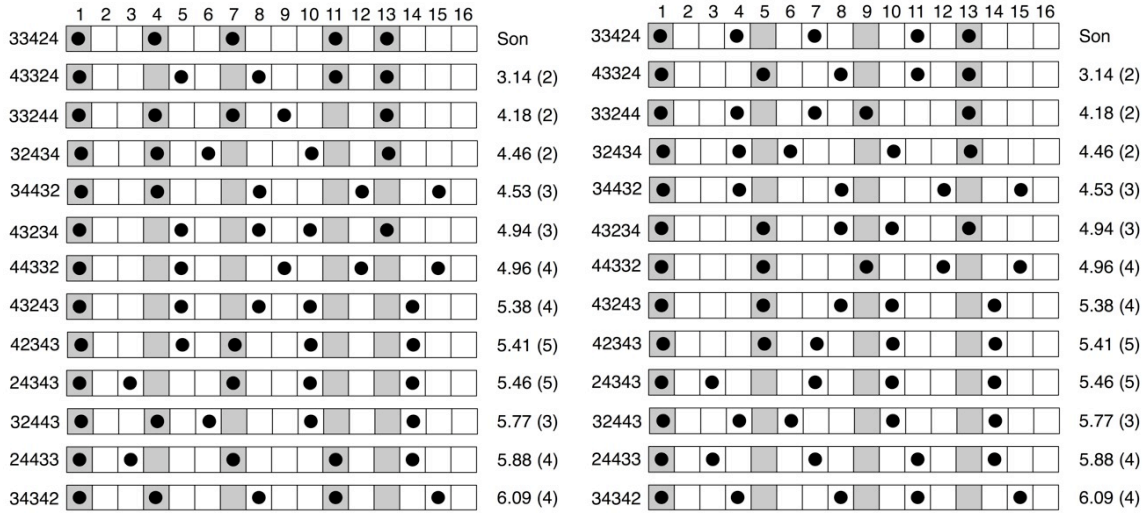

Fig. 15: The rhythmic permutations of the son ranked by increasing perceptual distance from the son, and the corresponding perceptual and edit distances to the son. In the left column the rhythm of the son is highlighted; in the right column the quarter-note pulses are highlighted.

The column on the right, in which the quarter-note beats are highlighted, brings a related pattern to light. Consider that the three most similar permutations to the son, all have an onset on the fourth beat (pulse 13). The next group of permutations, numbers 5-8, each have an onset on the second beat (pulse 5). Interestingly, the last four permutations, perceived as the most different from the son, do not have onsets on any quarter note pulse other than the downbeat. A possible hypothesis is that the less a rhythm sounded like the son, the more participants still equated it somehow with metric regularity. Perhaps this is not surprising, since the son is relatively easy to perceive in quadruple meter. (And again, historically it has proved to be appreciated and recognizable.) The rhythms furthest removed from the son, that do not share any onsets on any beat other than the downbeat, seem less likely to be perceived in 4/4 time, and also less likely to be 'catchy.'. It should be noted, however, that the perceptual distances between the last six rhythms are quite small; for example, between rhythms 7 and 9 the difference is only 0.08 on a scale from 1 to 9.

DISCUSSION

As the experiments presented above suggest, the edit distance model may well be a solid predictor of the similarity between rhythms as judged by human listeners. However, the model is far from infallible. A

discrepancy between the model and human perception was encountered in Experiment I, where changes at pulse 1 and 9 yielded highly salient changes, which the edit distance model did not reflect. From a metric perspective, the salience of pulses 1 and 9 is probably not surprising, since pulse 1 represents the downbeat in a 4/4 meter and pulse 9 the third beat, and both beats 1 and 3 are metrically strong in a 4/4 context. The weakness this suggests in the edit distance model is that it fails to account for metric hierarchies. Perhaps future research could indicate how the edit distance model can be adjusted for this omission.

A second discrepancy between the edit distance model and the human similarity judgments appeared in Experiment II. Consider the rhythms numbered 9 and 10 from Figure 15: 24343 and 32443. The former has an edit distance of 5 to the son and the second rhythm has an edit distance of 3 to the son, yet in the human distance ratings the first rhythm was rated lower than the second: 5.46 versus 5.77. While it may not be possible to fully account for this difference, it does encourage us to look more precisely at the edit distance computation. Recall that the edit distance is the minimum number of possible substitutions, deletions, and insertions required to convert one rhythm to the other, say, 32443 to 33424 (rhythm 10 to the son). Figure 16 illustrates the three operations that turn rhythm 10 into the son. The first operation performed on 32443 is the deletion of the fifth onset at pulse 14, yielding the 15-pulse, 4-onset rhythm 3246 (see Figure 15, second row). The second operation is the substitution of the silence at pulse 12 by an onset, resulting in the 15-pulse, 5-onset rhythm 32424 on the third row. The third and final operation is an insertion of a silence between pulses 5 and 6, yielding the desired son rhythm, 33424. Of these simple steps in terms of the edit distance model, both steps one and three could well result in substantial changes in terms of human perception. The deletion in step one reduces the length of the rhythm to 15 pulses and only 4 onsets. While most of the rhythm is still preserved, the 'feel' of the rhythm is probably changed considerably, as it has lost its quadruple character. (The exact salience of a deletion may of course also depend on the context and the method with which the salience is tested.) The opposite change is made in step 3, and again, intuitively it does not seem difficult to imagine situations in which an insertion in a 15/16 rhythm leads to a very salient change. Further tests should be conducted to study the salience of insertions, deletions, and substitutions specifically.

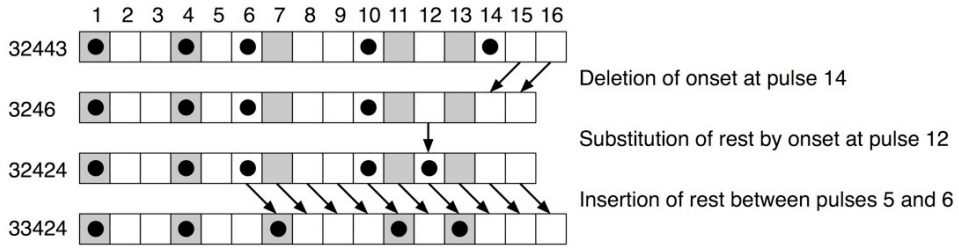

Fig. 16: The conversion of 32443 to 33424 (the son) according to the edit distance model.

On the other hand, the fact that the edit distance model allows for length changes in the rhythm can also be seen as an indicator of the strength of the model, as it obviously allows for comparisons between rhythms of different lengths. (This is a much more complicated matter for a model such as GTTM, where every time-signature has its own metric hierarchy.) This aspect of it can be particularly useful in studying sets of rhythms that are different and yet related, as found, for example, in African percussion music. Future experiments will most likely focus on testing the effectiveness of the edit distance model as a predictor of the similarity between rhythms of different lengths.

An unambiguous limitation of the edit distance model, as applied here, is that we could only apply the model to monophonic rhythms, i.e., a single stream of pulses. In many musical styles, one rarely encounters a single rhythm sounding at a given time. Another thought for future research, therefore, is to study possibilities for the application of the edit distance model in more challenging rhythmic settings.

ACKNOWLEDGMENTS

This research was funded primarily by the Mind, Brain, and Behavior Program at Harvard University. The second author was also financially supported by the National Sciences and Engineering Research Council of Canada (NSERC), administered through McGill University, Montreal, and by the Radcliffe Institute for Advanced Study at Harvard University, Cambridge, MA.

NOTES

[1] Toussaint (2002) and Huron and Ommen (2006) designed their measures of rhythm complexity and syncopation, respectively, on the basis of this hierarchy.

[2] The Mantel test computes the correlation coefficient between two dissimilarity matrices. It is especially designed to take into account the fact that entries in such matrices are not independent. It is based on making a large number of random permutations (we used 10,000) of the rows and columns of the matrices to obtain estimates of the distribution of standard correlations.

[3] As Toussaint has argued (Toussaint, 2011), the son has exhibited great appeal since its earliest documented historical appearance in 13th-century Bagdad.

REFERENCES

Bonnet, E., & Van de Peer, Y. (2002). zt: a software tool for simple and partial Mantel tests, *Journal of Statistical Software*, Vol. 7, No. 10, pp. 1-12.

Gascuel, O. (1997). *BIONJ*: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution,* Vol. 14, No. 7, pp. 685-695.

Hannon, E.E., & Trehub, S.E. (2005). Metrical categories in infancy and adulthood. *Psychological Science*, Vol. 16, No. 1, pp. 48-55.

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge: The MIT Press.

Huron, D., & Ommen, A. (2006). An empirical study of syncopation in American popular music, 1890-1939. *Music Theory Spectrum*, Vol. 28, No. 2, pp. 211-231.

Huson, D.H. (1998). *SplitsTree*: A program for analyzing and visualizing evolutionary data. *Bioinformatics*, Vol. 14, No. 10, pp. 68-73.

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge: The MIT Press.

Orpen, K.S., & Huron, D. (1992). Measurement of similarity in music: A quantitative approach for non-parametric representations. *Computers in Music Research*, Vol. 4, Fall, pp. 1-44.

Palmer, C., & Krumhansl, C.L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 16, No. 4, pp. 728-741.

Sankoff, D., & Kruskal, J. (1999). *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Stanford: CLSI Publications.

Scavone, G.P., Lakatos, S., & Harbke, C.R. (2002). The *Sonic Mapper:* An interactive program for obtaining similarity ratings with auditory stimuli. *Proceedings of the International Conference on Auditory Display* (pp. 1-4). Kyoto, Japan.

Sink, P.E. (1983). Effects of rhythmic and melodic alterations on rhythmic perception. *Journal of Research in Music Education*, Vol. 31, No. 2, pp. 101-113.

Sink, P.E. (1984). Effects of rhythmic and melodic alterations and selected musical experiences on rhythmic processing. *Journal of Research in Music Education*, Vol. 32, No. 3, pp. 177-193.

Smith, L.M. (2010). Rhythmic similarity using metrical profile matching. *Proceedings of the International Computer Music Conference*, New York, June 1-5, pp. 177-182.

Toussaint, G.T. (2011). The rhythm that conquered the world: What makes a "good" rhythm good? *Percussive Notes*. Towson, November 2011, in press.

Toussaint, G.T. (2004). A comparison of rhythmic similarity measures. *Proceedings of ISMIR 2004: 5th International Conference on Music Information Retrieval,* Universitat Pompeu Fabra, Barcelona, Spain, October 10-14, pp. 242-245.

Toussaint, G.T. (2002). A mathematical analysis of African, Brazilian, and Cuban clave rhythms. *BRIDGES: Mathematical Connections in Art, Music and Science*. Towson, Maryland: Towson University, pp. 157-168.